SAMI2021

IEEE 19th World Symposium on Applied Machine Intelligence and Informatics

Lucerne University of Applied Sciences and Arts

HOCHSCHULE LUZERN

Towards Granular Knowledge Structures: Comparison of Different Approaches

F. Stalder , A. Denzler , ***L. Mazzola***

Lucerne University of Applied Sciences and Arts (HSLU), Switzerland

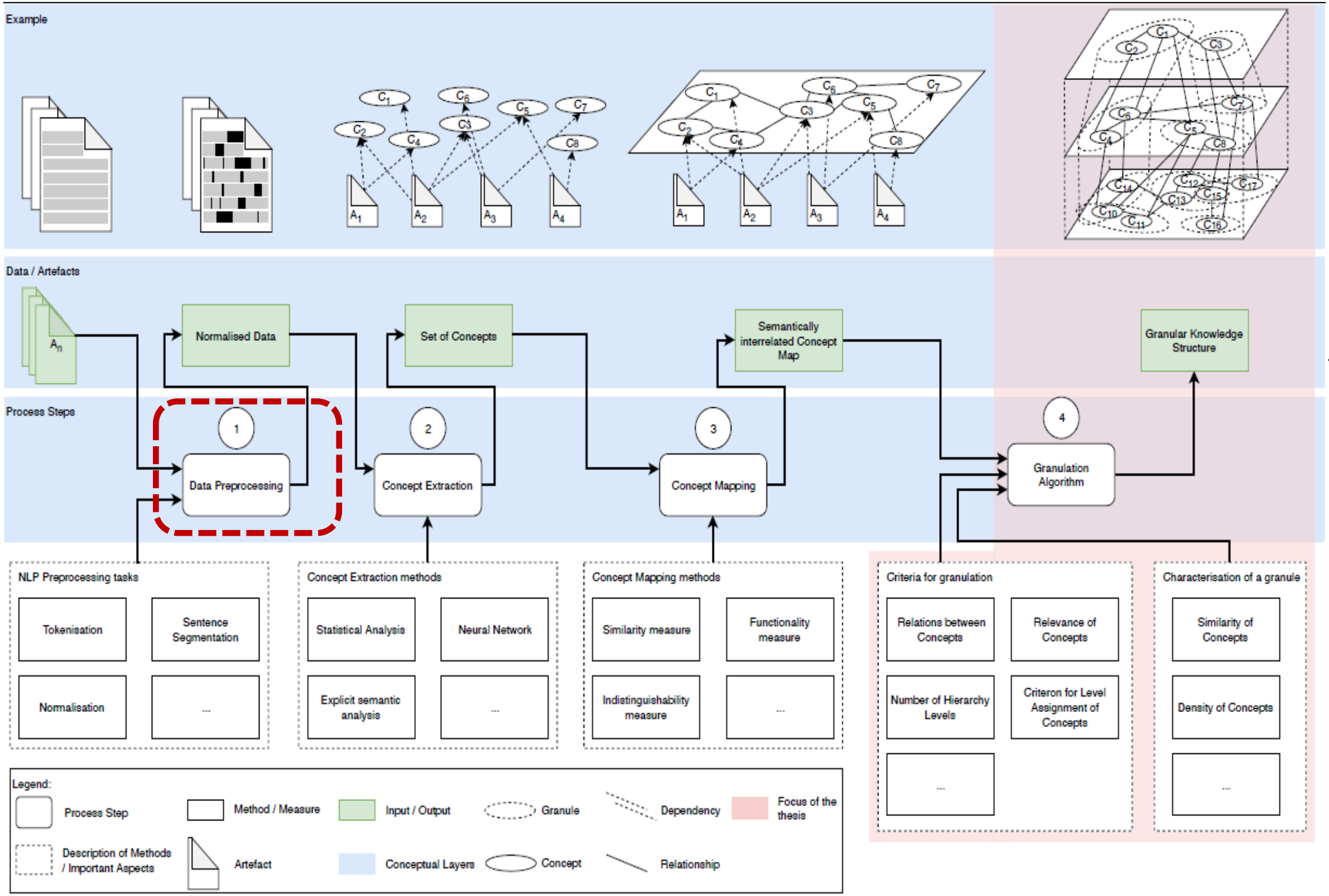
School of Computer Science

IEEE 19th World Symposium on Applied Machine Intelligence and Informatics

January 21-23, 2021, ~~in Herl'any, Slovakia~~ online

Intro

- Building granular knowledge structure (GKS) is a task becoming relevant
- Granular computing is only a reference model: it lacks specific algorithmic implementations as references
- We need to identify usable approaches for Granular Knowledge Map identification.



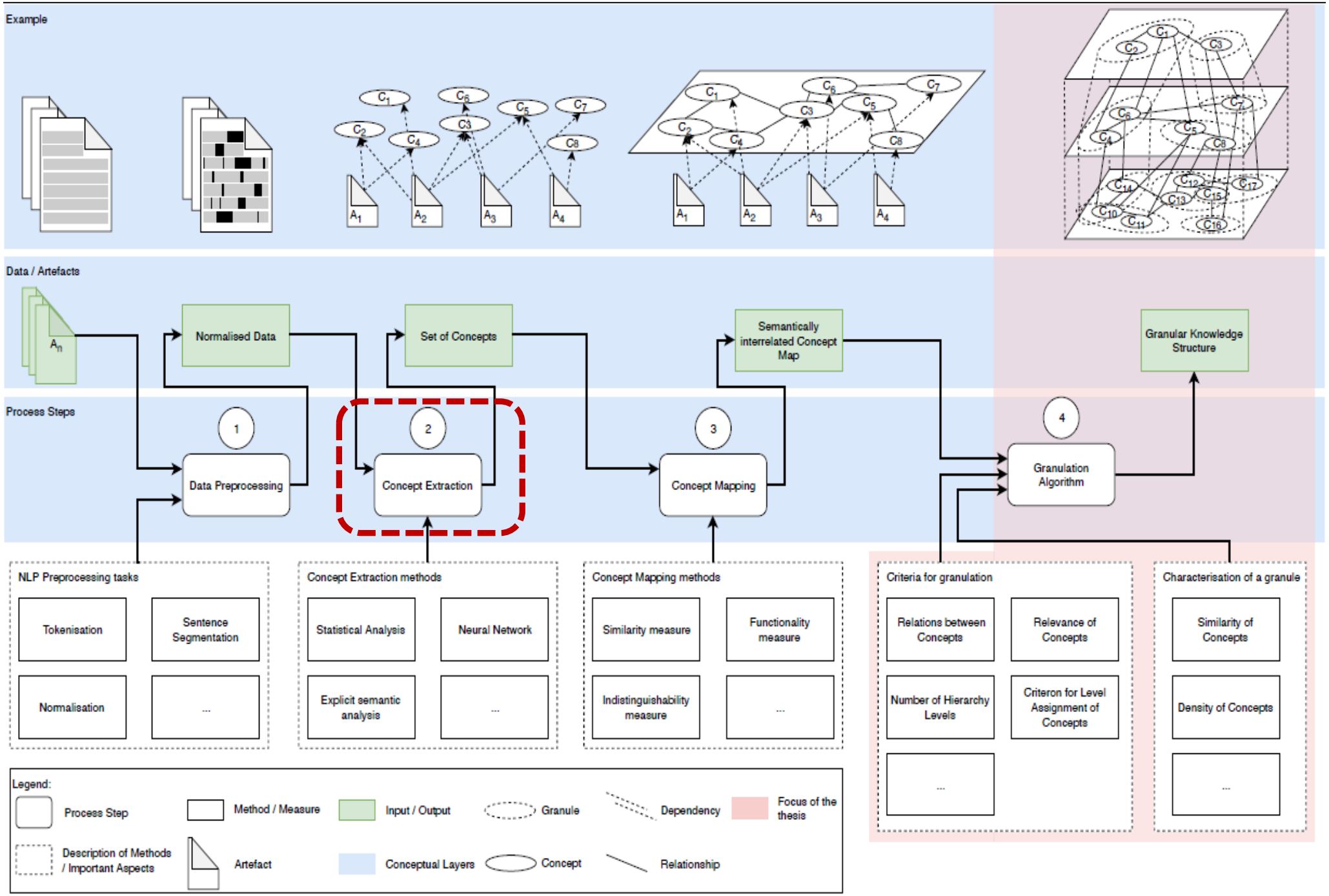
Lucerne University of Applied Sciences and Arts

HOCHSCHULE LUZERN

STEP 1:
Data Preprocessing

Objective:
Text normalization

Methods:
- Stop-words removal
- Tokenisation
- Normalisation
- (Sentence) segmentation
- …



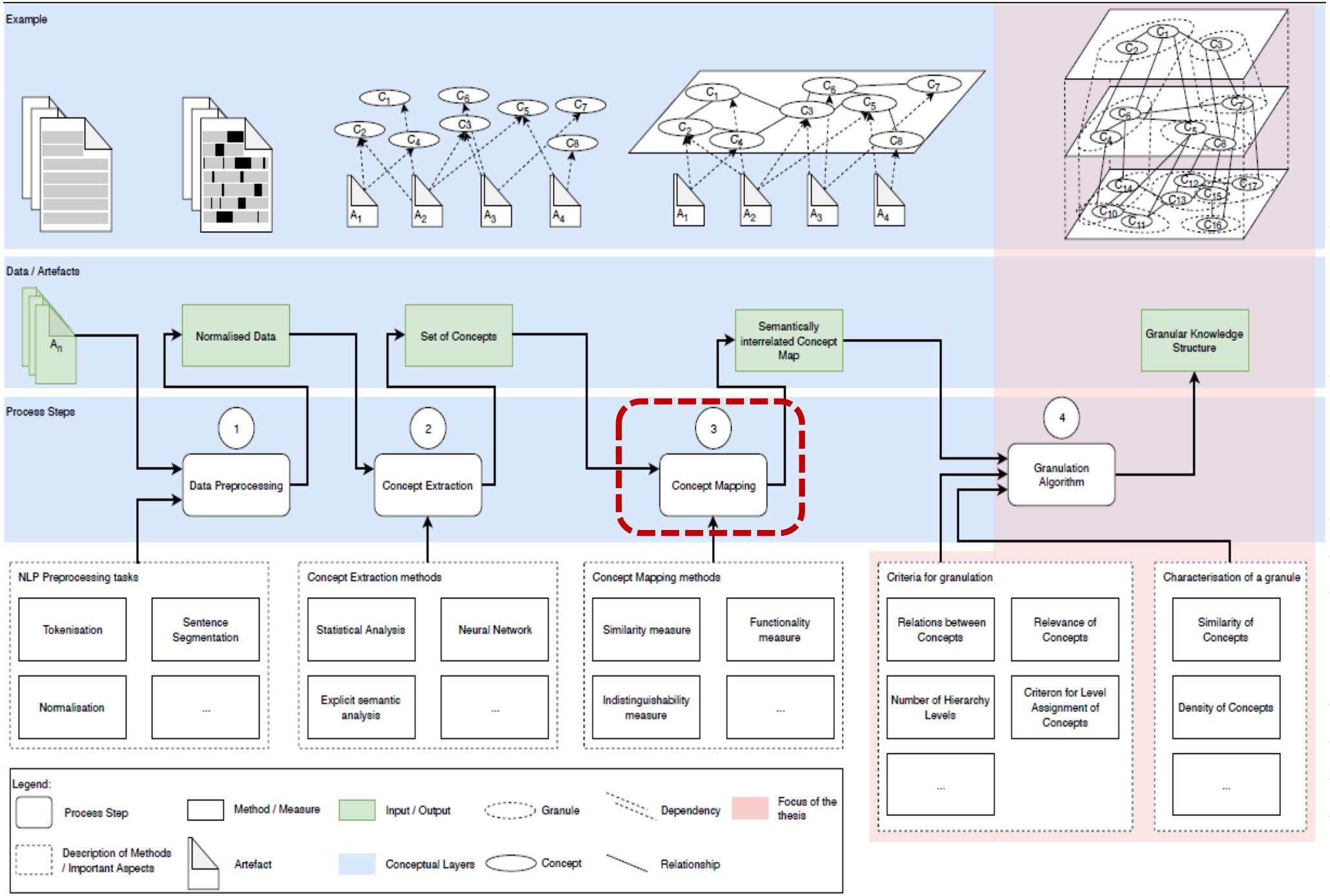
Lucerne University of Applied Sciences and Arts

HOCHSCHULE LUZERN

STEP 2:
Concept Extraction

Objective:
Set of Concepts

Methods:
- Statistical Analysis
- ESA
- NN
- Semantic role labeler
- ...



Lucerne University of Applied Sciences and Arts

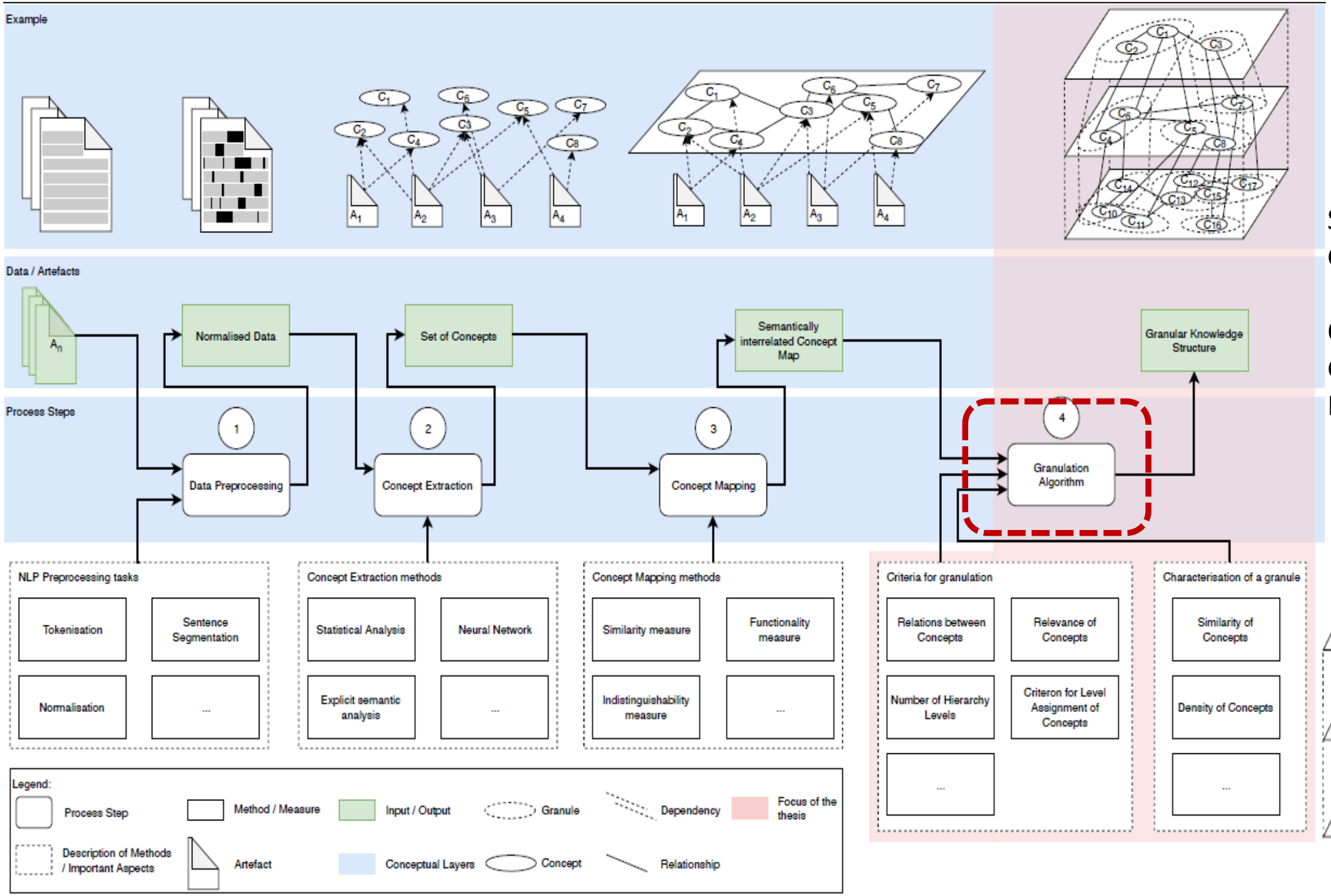
HOCHSCHULE LUZERN

STEP 3:
Concepts Mapping

Objective:
Semantic interrelation between concepts

Methods:
- Concepts linkage
- Document linkage
- …

Metrics:
- Similarity
- Indistinguishability
- Functionality
- …

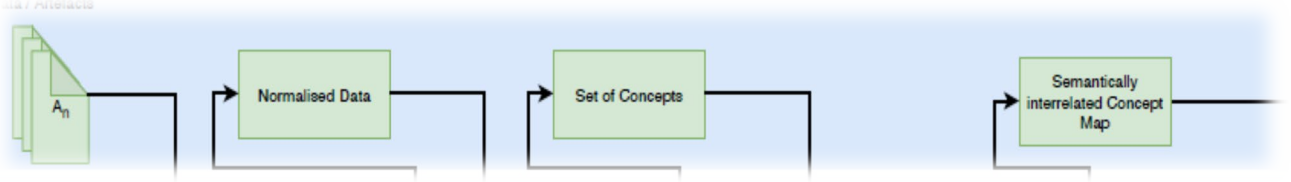


STEP 4 (the focus):
Granular approach

Objective:
Granular Knowledge Map

City
District
Street

Characteristics (steps 0-3)

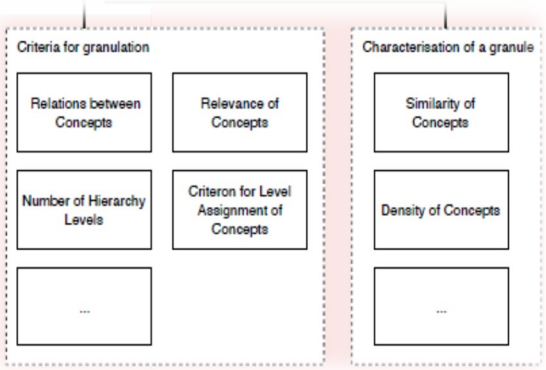


- Data: complete and noise-free
- Normalised data: domain-relevant and representative
- Set of Concepts: semantically meaningful and domain-relevant
  - Coverage of the full original data set
  - High specificity and precision (eg: TF_IDF), high TP and low FP
  - Coverage is of less importance (FN)
- Semantically interrelated Concept Map: dense-enough relations
  - Based on distance measure on high dimensional spaces
  - Need to find a meaningful cut-off/threshold value (to filter irrelevant relations)

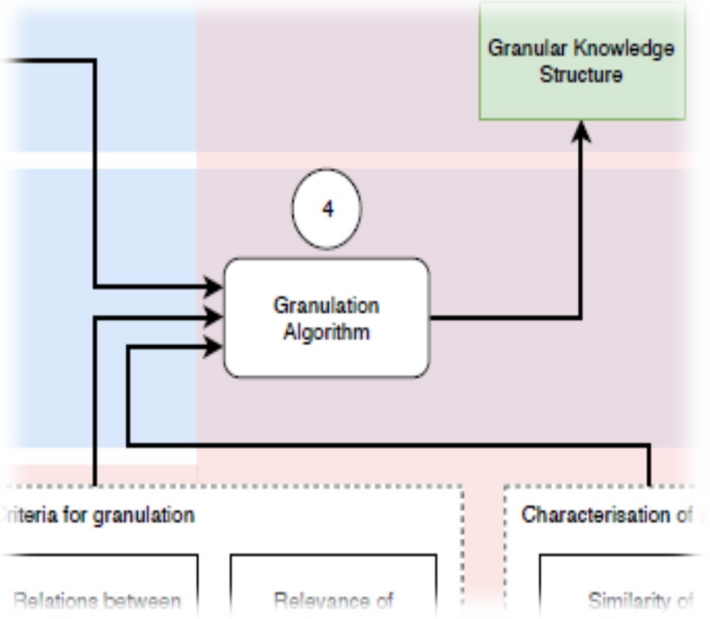
Characteristics (step 4: clustering)

- Clustering: defining homogeneous sets of concepts
  - Capability of manage fuzziness → concepts in different group with different confidences

- Hierarchy identification: bottom-up or top-down
  eg: granules dimension to decide the appropriate layer
  - Horizontal relationships (on a layer) can rely on averaged distance
  - Vertical relationships (within layers) can use averaged inter-granular relatedness measure
  - Can be improved by metadata (if available)



Evaluation dimensions



- Clustering Output :
  - Hierarchical vs. flat
  - Crisp-clustering vs. fuzzy groups boundaries (soft)
  - Same vs. variable dimension clusters
- Preliminary Input:
  - Cluster numbers, termination criteria, MAX cluster size, …
- Cluster Computation:
  - Based on Entity, Nodes, or Value Space
- Adaptive Learning (adapt underlying structure to changing conditions)
- Complexity (asymptotic estimation of time required for a solution)

Requirements (hard vs soft)

- HARD (should be addressed in the main algorithm):
  - Fuzzy clustering (crisp output is not enough)
- SOFT (can be achieved by combination with other algorithms):
  - Hierarchical output (with flat output, another algorithm should take care of inducing a hierarchy)
    - Eg: SOM with its GHSOM extension
  - Addressing high dimensional data (if appropriate)
- DESIRED:
  - Higher number of preliminary Inputs (hard to determine, but offer control over the algorithm)
  - Capacity of computing using different metrics
  - Adaptive learning, if possible both is unsupervised and supervised flavor
  - Lower possible time-complexity (to guarantee better scalability to larger dataset input)

		Partition based	Fuzzy Theory based	Agglomerative Hierarchical	Divisive Hierarchical	Distribution based	Density based	Graph Theory based	Grid based	Fractal Theory based	Self-organising Map	Projective
Characteristics	**Clustering Output**											
	Flat	x	x			x	x	x			x	x
	Hierarchical			x	x				x	x		
	Hard	x		x	x		x	x	x	x		x
	Soft		x			x					x	x
	Varying Cluster Size	x	x			x	x	x	x	x	x	x
	Preliminary Input											
	Number of Clusters	x	x									(x)
	Stopping Criterion	x	x								x	
	Cluster Size						x		x		x	(x)
	Special Parameters						x	x	x	x	x	x
	Computation of Clusters											
	Entities (Nodes)	x	x	x	x	x					x	x
	Relations (Edges)							x				
	Value Space						x		x	x		x
	Adaptive Learning											
	Supervised										x	x
	Unsupervised										x	x
AGGREGATED SCORE		**6**	**6**	**3**	**3**	**4**	**6**	**5**	**6**	**5**	**9**	**9** *(+2)*
Improvements	Hierarchical Clusters	x	x					x			x	x
	Soft Clusters			x			x	x	x			
	Adaptive Learning			x				x				
Complexity	Time Complexity(*asymptotic estimation*)	$\mathcal{O}(n*m*k*I)$	$\mathcal{O}(n*m*k^2*I)$	$\mathcal{O}(n^2*\log n)$	$\mathcal{O}(e*\log(v))$	$\mathcal{O}(n*\log n)$	$\mathcal{O}(n^2)$	$\mathcal{O}(e*d*\log v)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(Mn)$	$\mathcal{O}(n+k^2)$

Lucerne University of Applied Sciences and Arts

HOCHSCHULE LUZERN

Filtering by:
Fuzziness support
(required natively)

		Partition based	Fuzzy Theory based	Agglomerative Hierarchical	Divisive Hierarchical	Distribution based	Density based	Graph Theory based	Grid based	Fractal Theory based	Self-organising Map	Projective
Characteristics	**Clustering Output**											
	Flat	x	x			x	x	x			x	x
	Hierarchical			x	x				x	x		
	Hard	x		x	x		x	x	x	x		x
	Soft		x			x					x	x
	Varying Cluster Size	x	x			x	x	x	x	x	x	x
	Preliminary Input											
	Number of Clusters	x	x									(x)
	Stopping Criterion	x	x								x	
	Cluster Size						x		x		x	(x)
	Special Parameters						x	x	x	x	x	x
	Computation of Clusters											
	Entities (Nodes)	x	x	x	x	x					x	x
	Relations (Edges)							x				
	Value Space						x		x	x		x
	Adaptive Learning											
	Supervised										x	x
	Unsupervised										x	x
AGGREGATED SCORE		**6**	**6**	**3**	**3**	**4**	**6**	**5**	**6**	**5**	**9**	**9** *(+2)*
Improvements	Hierarchical Clusters	x	x					x			x	x
	Soft Clusters			x			x	x	x			
	Adaptive Learning			x				x				
Complexity	Time Complexity(*asymptotic estimation*)	$\mathcal{O}(n*m*k*I)$	$\mathcal{O}(n*m*k^2*I)$	$\mathcal{O}(n^2*\log n)$	$\mathcal{O}(e*\log(v))$	$\mathcal{O}(n*\log n)$	$\mathcal{O}(n^2)$	$\mathcal{O}(e*d*\log v)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(Mn)$	$\mathcal{O}(n+k^2)$

Lucerne University of Applied Sciences and Arts

HOCHSCHULE LUZERN

Filtering by:
Hierarchical structure support
(either native or in an extension, existing or not)

		Partition based	Fuzzy Theory based	Agglomerative Hierarchical	Divisive Hierarchical	Distribution based	Density based	Graph Theory based	Grid based	Fractal Theory based	Self-organising Map	Projective
Characteristics	**Clustering Output**											
	Flat	x	x			x	x	x			x	x
	Hierarchical			x	x				x	x		
	Hard	x		x	x		x	x	x	x		x
	Soft		x			x					x	x
	Varying Cluster Size	x	x			x	x	x	x	x	x	x
	Preliminary Input											
	Number of Clusters	x	x									(x)
	Stopping Criterion	x	x								x	
	Cluster Size						x		x		x	(x)
	Special Parameters						x	x	x	x	x	x
	Computation of Clusters											
	Entities (Nodes)	x	x	x	x	x					x	x
	Relations (Edges)							x				
	Value Space						x		x	x		x
	Adaptive Learning											
	Supervised										x	x
	Unsupervised										x	x
AGGREGATED SCORE		6	6	3	3	4	6	5	6	5	9	9 *(+2)*
Improvements	Hierarchical Clusters	x	x					x			x	x
	Soft Clusters			x			x	x	x			
	Adaptive Learning			x				x				
Complexity	Time Complexity(*asymptotic estimation*)	$\mathcal{O}(n*m*k*I)$	$\mathcal{O}(n*m*k^2*I)$	$\mathcal{O}(n^2*\log n)$	$\mathcal{O}(e*\log(v))$	$\mathcal{O}(n*\log n)$	$\mathcal{O}(n^2)$	$\mathcal{O}(e*d*\log v)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(Mn)$	$\mathcal{O}(n+k^2)$

Lucerne University of Applied Sciences and Arts

HOCHSCHULE LUZERN

Comparing on the remaining (desiderata)

Two candidates:
- GHSOM
- Projective clustering (eg: ensembles)

Complexity:
- SOM α product of data points (n) with number of neurons (M) in the lattice
- Ensembles α data points (n) (cluster number is normally significantly smaller)

GHSOM

??????

Results

- We discussed the process to generate Granular Knowledge Maps, based on its 4 basic steps
- For each step, we described possible methods and requirements of the input data/artefact
- Concentrating on the clustering and hierarchy building, we compared 11 families of algorithms and discover the best two candidates, based on their asymptotic computational (time) complexity:
  - Low dimensionality data: Growing Hierarchical Self-Organizing Maps (GHSOM)
  - High dimensionality: a projective approach, such as Projective clustering ensembles
    - hierarchical extension should be added on top of it

Conclusion and outlook

- Findings:
  - Absence of an universal solution
  - 2 candidates ranked best for granular knowledge structure (GKS) creation
    - One for low-dimensional space, the other for high dimensional ones
  - Extension should be added to fulfill all the requirements identified
- What's next?
  - theoretical work, need validation by external measures (eg: expert feedback)
  - Compare the performance on different datasets, for generalization purposes
  - Explore acceptance of such a solution
    - by collecting feedback from user for semantic meaningfulness
    - By rating the results produced using the GKS as knowledge base

Lucerne University of
Applied Sciences and Arts
HOCHSCHULE
LUZERN

Questions?

- For any questions or request, please feel free to contact us, thanks.

- By email:
  - Florian Stalder florian.stalder@hslu.ch
  - Alexander Denzler alexander.denzler@hslu.ch
  - Luca Mazzola luca.mazzola@hslu.ch or mazzola.luca@gmail.com

- If you are interested in our activities, please visit our lab website:
  - http://hslu.ch/blockchainlab/

SAMI2021

IEEE 19th World Symposium on Applied Machine Intelligence and Informatics